Comparative Study of Probability Models for Compound Similarity Searching

Naomie Salim
Faculty of Computer Science
and Information Systems
Universiti Teknologi Malaysia
naomie@fsksm.utm.my

Mercy Trinovianti Mulyadi
Program Pengajian Diploma,
Universiti Teknologi Malaysia
mercy@citycampus.utm.my

Abstract - *The quality of a chemical retrieval system heavily depends on its molecular similarity function which returns a similarity measurement between the target compound and each molecule in the collection. Compounds are sorted according to their similarity values with the query and those with high ranks are returned to the users. Most current chemical retrieval systems use the vector space model for similarity calculation. In this paper, the use of probability of relevance for compound retrieval is explored. It reports on the effectiveness of the probability model for compound similarity searching by using Binary Independence Model and Binary Dependence Model on two different databases. The result based on fusion of queries for both models is also discussed. The results show that in all cases, Binary Independence Retrieval model performed better than Binary Dependence model. It is also found that fusion does not give better results than the un-fused queries.*

Keywords: molecular similarity searching, probability model, data fusion

1 Introduction

Calculation of similarity between two molecules is an important tool in a chemoinformatics software. A common application of similarity searching is in the rational design of new drugs and pesticides where the nearest neighbours for an initial lead compound are sought in order to find better compounds. Similarity searching is also used for property prediction purposes where the properties of an unknown compound are estimated from those of its nearest neighbours. Underpinning these applications of molecular similarity measure is the similar property principle which states that structurally similar molecules will exhibit similar physiochemical and biological properties. Related to the similar property principle is the concept of neighbourhood behavior which states that compounds within the same neighbourhood or similarity region have the same activity. Unknown biological or physicochemical properties of a molecule can be predicted from the properties of molecules that lie within the same neighbourhood region. In lead finding, selection of compounds whose neighbourhood regions overlap one another should be avoided. In lead optimisation, if a particular compound is found to be active, compounds that lie in the same neighbourhood region can be tested to find one with the most optimum activity.

Molecular similarity searching generally consists of two steps. The first step is to define a representation for the molecules. The molecules can be represented in one-dimensional, two-dimensional and three-dimensional representations. The second step is to compare the selected representations of two or more molecules by calculating similarity coefficient to assign a similarity index. Chemical similarity searching using 2D molecular representations or bit-strings with the Tanimoto similarity coefficient seems to work reasonable well in terms of both efficiency and effectiveness [1]. Algorithms developed for the processing of textual database are also applicable to the processing of chemical structure database [2]. In this paper, another alternative for compound similarity searching based on Probability Model is proposed. Probability Model ranks chemical compounds in decreasing order of their probability of being similarly active to the target compound. According to the Probability Ranking Principle (PRP), if the ranking of the compounds is in decreasing probability of usefulness to the user, then the overall effectiveness of the system to its users will be the best [3].

As an introduction, we will briefly explain the concept of the two Probability Models (PM) used: the Binary Independence Retrieval Model (BIR) and the Binary Dependence Model (BDM). Then, we will show the results of a pilot experiment which that applied PM for chemical compound retrieval. In section 5 we will elaborate the results of experiments on a different database with a wider range of activities.

2 Binary Independence Model for Chemical Compound Retrieval

The simplest of the PM models is based on the presence or absence of independently distributed bits in active and inactive structures, i.e. the distribution of any given bit over the collection of structures is assumed to be independent of the distribution of any other bit. The probability of any given bit occurring in an active structures is independent of the probability of any other bit occurring in an active structure (and similarly for inactive structures). The model is referred to as the

"binary independent"(BI) model. This model also known as the relevance weighting theory in the more general literature of information retrieval [4]. We start with p_i, the probability that bit b_i appears in a structure given that the structure is active and q_i, the probability that b_i appears in a structure given that the structure is inactive. Using Bayes' rule of inference and the BI assumption, one can derive a function for ranking structures by probability of relevance. In this function, each bit *bi* receives a weight *wi* given by:

$$w_i = \log \frac{p_i \cdot (1 - q_i)}{q_i \cdot (1 - p_i)}$$

The "odds" of b_i appearing in a active structure is $p_i/(1-p_i)$. Similarly, the "odds" of b_i appearing in a inactive structures is $q_i/(1-q_i)$. Hence, w_i measures the odds of b_i appearing in a active structure divided by the odds of b_i appearing in a inactive structure, i.e., the odds ratio. Taking the log makes this function symmetric: w_i=0 if p_i=q_i, is positive if p_i>q_i, and is negative if p_i<q_i. w_i is a good measure of how well a bit can distinguish active from inactive structures. The function w_i is called the "term relevance" weight or "term relevance function" or "logodds," i.e., the log of the odds ratio for b_i. Plainly, w_i will be a very large positive number for a bit that has a high probability of appearing in a active structure and very low probability of appearing in a inactive structure (and a very large negative number if the probabilities are reserved). Moreover, if the set of index bits in a structure collection satisfies the BI condition, the odd ratio (odds of active divided by odds of non-active) for a given structure *S* is merely the odds ratios of all the *bi* bits appearing in *S*, i.e, the product of all the corresponding w_i. Hence the log of the odds ratio for *S* can be computed as the sum of the w_i. In other words, the logodds of *S* being active is computed as the sum of the w_i for index bits appearing in *S*. The trick is to find a way of computing p_i and q_i. If active data is available, e.g from manually evaluating a previous run with the same query against the same collection or against a training set, then p_i and q_i can be estimated from a 2x2 "contingency" Table 1 summarizing the relevance judgments as given below:

	No. of Active Structures	No. of Inactive Structures	Total
No. of structures including bit b_i	*a*	*n-a*	*n*
No. of structures excluding bit b_i	*A-a*	*(N-R)-(n-r)*	*N-n*
Total	*A*	*N-A*	*N*

Table 1. Contingency table of relevance judgments

Here *N* is the total of structures in the collection, *n* is the total number of structures that contain bit b_i, *A* is the total number of active structures retrieved, and *a* is the total number of active structures retrieved that contain bit b_i. From this table, we can estimate p_i as *r/R* (the proportion of active structures containing t_i), and q_i as *(n-a)(N-A)* (the proportion of inactive structures containing b_i). Obviously, this assumes that "the bit distribution in the active items previously retrieved is the same as the distribution for the complete set of active items, and that all non-retrieved items *[N-A]* can be treated as inactive"[5]. The latter assumption is necessary to allow us to assume that *N-A* (all retrieved inactive structures plus all non-retrieved structures) equals the total number of inactive structures. The former assumption allows us to treat the proportion of active structures containing b_i in the retrieved sample as characteristic of the proportion in the complete collection, for all b_i.

Equivalently, the odds of b_i appearing in an active structure is *(a/A)/(1-a/A)= a/(A-a)*, and the odds of b_i appearing in inactive structure is
(n-a)/(N-A)/[1-(n-a)/(N-A)]=(n-a)/(N-A-n+a)
Inserting these values into the formula for w_i, we obtain:

$$w_i = \frac{a(N - A - n + a)}{(A - a)(n - a)}$$

This formula for w_i obviously breaks down if p_i equals one (*a=A*) or zero (*a=0*). Similarly, w_i breaks down if q_i equals to one (*n-a=N-A*) or zero (*n=a*). Statistical theory has been used to justify modifying the formulas for p_i and q_i to avoid these singularities by adding a constant *c* to the numerator and one to the denominator, where *c*=0.5 or *c=n/N* [6]. However, there are always cases where these constants dominate the computation and distort the results. Shaw [7] proposes to avoid these problems by using the unmodified formulas for all cases except for singularities where alternative formulas are specified. If the constant 0.5 is inserted in the formula for w_i ,the result is:

$$w_i = \frac{(a + 0.5)(N - A + a + 0.5)}{(A - a + 0.5)(n - a + 0.5)}$$

It should be noted that in the bit relevance model described above, the probabilities of relevance and non relevance, given b_i and the corresponding logodds function w_i, are based entirely on the presence or absence of each bit in active and inactive structures. A bit b_i is favored, i.e given a high w_i, if it appears more frequently in active structures than in inactive structures. It however received no "extra credit" for appearing more frequently than another bit b_i in active structures.

3 Binary Dependence Model for Chemical Compound Retrieval

Bit dependencies refer to the presence or absence of a bit, which provides information about the probability of presence, or absence of another bit. Assume vector structure, $S = \{b_1, b_2 \ldots b_n\}$ are binary values. It is arbitrarily complex to capture all dependence data as we

need to condition each variable in turn on a steadily increasing set of other variables. Hence, to estimate the probability of a structure ($P(S)$) this model captures only the significant pairwise dependence information. Thus $P(S)$ is the probability of a bit i being solely dependent on some preceding bit $b_{j(i)}$:

$$P(S) = \prod_{i=1}^{n} P(b_i \mid b_{j(i)}) \qquad 0 \le j(i) \le i$$

A probability distribution that can be represented as in the above expression is called a probability distribution of first-order tree dependence [8], which suggest the construction of a *Maximum Spanning Tree* (MST) using the *Expected Mutual Information Measure* (EMIM). EMIM is a measure of a variable containing the information about another variable. Hence, it requires the counting of co-occurrences of bits in a structure, and thus used to measure the dependence between a pair of bits. Let $G(V,E)$ be a connected graph, where V is the set of nodes and E is the set of edges. Assigned to each edge $(i, j(i))$ is a weight $w_{(i,\ j(i))}$, obtained from calculating the EMIM value of the pair of variables. An MST is a tree that includes every node and has maximal total weight. It simply maximizes the sum:

$$\sum_{i,j} I(b_i, b_{j(i)})$$

where $I(b_i, b_{j(i)})$ represents the expected mutual information between bit b_i and $b_{j(i)}$,

$$I(b_i, b_{j(i)}) = \sum_{b_i, b_{j(i)}} P(b_i, b_{j(i)}) \log \frac{P(b_i, b_{j(i)})}{P(b_i) P(b_{j(i)})}$$

The contingency table (see Table 2) below further simplify the calculation of EMIM into the following:

$$I(b_i,b_{j(i)}) = (1)\log\frac{(1)}{(5)(7)} + (2)\log\frac{(2)}{(6)(7)} + (3)\log\frac{(3)}{(5)(8)} + (4)\log\frac{(4)}{(6)(8)}$$

	$b_i = 1$	$b_i = 0$	
$b_{j(i)} = 1$	(1)	(2)	(7)
$b_{j(i)} = 0$	(3)	(4)	(8)
	(5)	(6)	(9)

Table 2. Contingency table of maximum likelihood estimates

Hence, the first step in this model is to generate the MST to identify the most important pairwise dependencies. Each given chemical structure collection will construct an MST based on all bits included in the collection. There are many algorithms in generating an MST from pairwise association measures. We used the algorithm by Whitney (1972) which is based on the *Dijkstra* technique where a maximum spanning tree is grown by successively adjoining the farthest remaining node to a partially formed tree until all node of the graph are included in the tree. Next, the dependence tree is then used to expand the query by taking the original query bits and adding all bits that are immediately adjacent in the MST. The pairwise term dependencies is obtained for all bit pairs b_i and b_j in the expanded query such that each pair (b_i, b_j) is represented by an edge in the spanning tree. The following are the similarity function or RSV of this model, considering only the terms $P(S|A)$ and $P(S|NA)$. The rest remains as a constant and thus are not included in the calculation of RSV:

$$\frac{P(A|S)}{P(NA|S)} \propto \sum_{i=1}^{n} [b_i \log \frac{p_i(1-q_i)}{q_i(1-p_i)}] + \sum_{i=2}^{n} b_{j(i)} [\log \frac{p_{j(i)} - p_{i|j(i)}}{p_{j(i)}(1-p_i)} - \log \frac{q_{j(i)} - q_{i|j(i)}}{q_{j(i)}(1-q_i)}]$$
$$+ \sum_{i=2}^{n} b_i b_{j(i)} [\log \frac{p_{i|j(i)}(1-q_{i|j(i)})}{q_{i|j(i)}(1-p_{i|j(i)})} - \log \frac{p_i(1-q_i)}{q_i(1-p_i)} - \log \frac{p_{j(i)}(1-q_{j(i)})}{q_{j(i)}(1-p_{j(i)})}]$$

Relevance (activity) information can be considered to be available upon experimentation. To test the algorithm, we used information on the activity of compounds that is available in the database. This model computes the probability of $P(S|A)$ and $P(S|NA)$ using the same contingency table as the BIR model (Table 1) and thus producing the following assumption. The adjustment factor is also taken in to consideration to avoid problem resulting from the small value of A and a_i.

a) $p_i = (a_i + 0.5) / (A + 1)$
b) $q_i = (n_i - a_i + 0.5) / (N - A + 1)$
c) $p_{j(i)} = (a_{j(i)} + 0.5) / (A + 1)$
d) $q_{j(i)} = (n_{j(i)} - a_{j(i)} + 0.5) / (N - A + 1)$
e) $p_{i|j(i)} = (a_{i|j(i)} + 0.5) / (A + 1)$
f) $q_{i|j(i)} = (n_{i|j(i)} - a_{i|j(i)} + 0.5) / (N - A + 1)$

where

N-is the number of structures in database
n_i-refers to the frequency of structure containing bit b_i
$n_{j(i)}$-refers the frequency of structure containing bit $b_{j(i)}$
$n_{i|j(i)}$-refers to the frequency of structure containing both bit b_i and bit $b_{j(i)}$
A-is the total number of active structures
a-refers to the total number of active structures containing a particular bit b
b_i-refers to bit b at location i,
$b_{j(i)}$-refers to bit b at location $j(i)$ where bit $b_{j(i)}$ is the preceding bit of bit b_i,
$p_{i|j(i)}$ -is the probability of both bit b_i and bit $b_{j(i)}$ appearing in active structures,
p_i-is the probability of both bit b_i appearing in active structures,
$q_{i|j(i)}$-is the probability of both bit b_i and bit $b_{j(i)}$ appearing in inactive structures,
q_i-is the probability of both bit b_i appearing in inactive structures,

4 Probability Model on the AIDS Dataset

The objective of the research was to apply the Probability Model in chemical compound similarity searching [9] to investigate whether the proposed approach yields better performance of screening chemical compounds compared to existing methods. The experiment uses the National Cancer Institute (NCI) AIDS Database (National Cancer Institute, 1999) as the

test data set. It represents a large data set composed of both active and inactive compounds against a specific therapeutic target and provides a thorough sampling of a particular region in chemical space. This public database contains 5772 compounds, including 247 confirmed active (CA), 802 confirmed moderately active (CM) and 4723 confirmed inactive (CI). In this data set, both the CA and CM are treated equally as actives and the CI as inactives. For the molecular representation, the BCI 1052-bit structural key-based bit string is used. It is generated based on the presence or absence of fragments in the BCI's standard 1052 fragments-dictionary. Probability model that are used are the Binary Independence Retrieval (BIR) and Binary Dependence (BD) Model. Two sets of experiment were conducted; the first experiment focuses on comparing the effectiveness of the proposed approaches in similarity searching. A series of simulated similarity searching is conducted for both proposed approaches. The second experiment investigates whether fusion of the query of the proposed probability models is better than the unfused query. Data fusion is an approach where data, evidence, or decisions coming from or based on multiple sources, about the same set of objects are integrated to increase the quality of decision making under uncertainty about the objects [10]. Three approaches in evaluating the performance of the search methods are used, mainly the GH score, initial enhancement and the number of actives at the top 5% of ranked list.

The GH score gives an indication of how good the retrieved list is with respect to a compromise rate of maximum yield and maximum percent of actives retrieved.

$$GH = \frac{H_a(A + H_t)}{2AH_t}$$

where A is the number of actives structures in the database,
H_t is the number of structures in a retrieved list, and
H_a is the number of active structures in a retrieved list.

a) Initial enhancements refers to a number of chemical structure retrieved before half of the actives are found. The lesser the value, better is the performance of the similarity searching system.

b) A high number of active structures at the top 5% of this list denotes a good similarity searching system.

For the first experiment, in the data set, each active compound acts as a probe to search the remainder of the data set. Compounds are then ranked according to the calculated similarity values, from most to least similar. Lastly, the ranked list is analyzed to determine which method is more effective. Mean while for the second experiment the NCI AIDS dataset is divided equally to four sets, with 1443 structures in each set. The NCI AIDS dataset organises compounds according to the following: CA, CM and CI. Hence, this simplifies the division of the data sets with each set having equal distribution of CA, CM and CI. Next, an active compound is posted as query. The similarity searching is conducted on the first set and it returns the top 100 compounds. Based on these compounds, the probability of p_i and q_i for each bit i is computed. It will then be used to obtained the ranking score function (RSV) for the second set. The same procedure is repeated again, where the probability of p_i and q_i obtained from the top 100 compounds of the second set is used to compute the RSV for the third set. Finally, the probability of p_i and q_i obtained from the top 100 compounds of the third set is used to compute the RSV for the fourth and final set. Thus, the result of each query posted will return a total number of 400 compounds obtained by combining the result of each set.

The result readings of GH scores are taken on each 5% interval of structure retrieved, which are 5%, 10%, 15% ...30%. Readings stop when 30% of the structures retrieved. This means that the final GH score value is obtained when 1732 compounds are retrieved. Retrieval of more compounds than this value is not considered because users normally will only look at the top 1000 compounds. Table 3 and 4 shows the average result of BIR and BD based similarity searching. Generally, the BD has a slightly better performance than BIR and the fused query actually yielded poorer results compared to un-fused queries. For the number of chemical structures retrieved before half of the actives are found, the BIR-based similarity searching attains an average of 1917 compounds and BD with 1859 compounds in the unfused experiment and 2188 compounds for BIR and 2236 for BD in the fusion experiment. In terms of number of active compounds at the top 5% of the ranked list, again the BD has a slightly better performance than BIR and the fused query again yielded poorer results compared to un-fused queries. It is thus concluded that BD model has better performance than the BIR model. Meanwhile for fusion experiments, both BIR and BD model performs poorly compared to un-fused experiments.

	5%	10%	15%	20%	25%	30%
BIR (no query fusion)	29.43	29.51	31.06	32.97	35.14	37.27
BD (no query fusion)	31.2	30.86	32.26	34.19	36.18	38.16
BIR (query fusion)	26.77	26.22	27.45	29.38	31.45	33.61
BD (query Fusion)	28.71	26.9	27.61	29.26	31.13	33.12

Table 3. Result analysis of average GH score for the AIDS database.

	Initial enhancement	Average number of activities at top 5%
BIR (no query fusion)	1917	133
BD (no query fusion)	1859	141
BIR (query fusion)	2188	120
BD (query Fusion)	2236	129

Table 4. Result analysis of average initial enhancement and average number of actives at top 5% for AIDS database.

5 Probability Model on the MDDR Dataset

The second batch of experiments were conducted on the MDDR database, where there are 799 different activities associated with the 113842 compounds in the database. In the MDDR database, all these activities are classified under several broad activity classes. A subset of 4448 compounds for testing three activities- Benzodiazepine Agonist, HIV Protease Inhibitor and ACE Inhibitor were built, as shown in Table 5.

Activity	Class	Compound
Benzodiazepine Agonist	Anxiolytics	257
HIV Protease Inhibitor	Agents for AIDS	926
ACE Inhibitor	Antihypertensive Agents	500
Others		2765
Total		**4448**

Table 5. Activities and compounds selected from the MDDR database

For testing of the three different active types, all other compounds is ensured to not have the activity tested in order to be treated as inactives. All the data is represented by the BCI bit strings, but the BCI 1056-bit structural key is used instead of the BCI 1052 bit, to make sure that the choice of bit strings does not have any effect on the results. The same methodology as described in section 4 is repeated. All the results shown are averaged over all the targets for the three different types of actives.

5.1 Experiment 1: Comparing the Effectiveness of Similarity Searching Methods

In the following figure(see Figure 1), GH score values at each 5% level of the structure retrieved are plotted for both BIR and BD models. As can be seen in the graph, we can conclude that BIR model performed significantly better than the BD model.

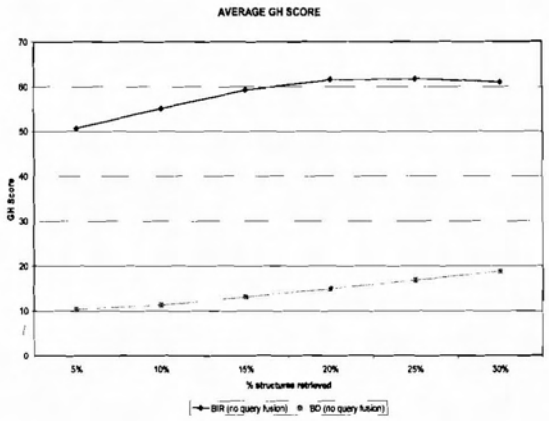

Figure 1. GH score analysis for BIR and BD model.

Initial enhancement refers to the number of chemical structures retrieved before half of the actives are found. Figure 2 below shows that again the BD has the highest value of initial enhancement. It retrieves approximately 2051 more structures before half of the activities are found compared to the BIR model. Here again we can see that BIR model is more superior in terms of average initial enhancement.

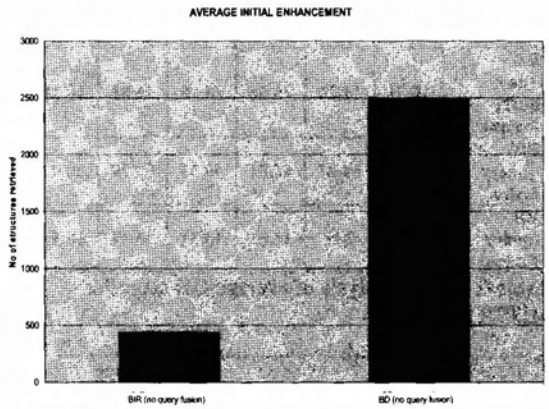

Figure 2. Initial enhancement analysis for BIR and BD model.

A good similarity searching system is also denoted by the number of active structures at the top 5% of its list. Top structures of the list normally show the nearest neighbours of the target molecule. More actives on this part of ranked list can help in the lead optimisation process, where initial lead compound are sought in order to find better compounds. The following Figure 3 shows us that the BIR has given more actives than the BD model. It also suggests that the BIR model is a more

effective similarity searching capability than the BD model.

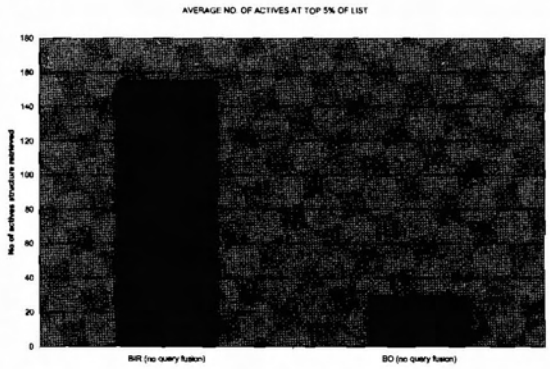

Figure 3. Average no of actives at top 5% for BIR and BD model.

5.2 Experiment 2: Comparing the Performance of Query Fusion Result of Similarity Searching Methods

In Table 6, it can be seen that in the fusion experiment, the BIR model still exceeds the BD model. Again the results suggest that query fusion here gives poorer results compared to unfused query.

	5%	10%	15%	20%	25%	30%
BIR (no query fusion	50.70	55.16	59.27	61.57	61.79	61.07
BD (no query fusion)	10.31	11.26	13.08	14.88	16.81	18.74
BIR (query fusion)	42.22	54.87	58.81	59.53	59.79	60.02
BD (query Fusion)	8.72	9.21	11.63	13.31	15.32	16.83

Table 6. Result analysis of average GH score for MDDR database

	Initial enhancement	Average number of actives at top 5%
BIR (no query fusion)(MDDR)	438	155
BD (no query fusion)(MDDR)	2489	29
BIR (query fusion)(MDDR)	509	143
BD(query Fusion)(MDDR)	3021	23

Table 7. Result analysis of Initial enhancement and Average number of actives at top 5% for MDDR database

Table 7 shows the average initial enhancement and number of actives in the top 5% compounds retrieved for two models in both fused and unfused experiments. Again it can be seen that in the query fusion gives poorer performance compared to unfused query.

6 Conclusion

The results of the experiments show that in many cases, the BIR model is superior to the BD model. However, for the AIDS dataset, the BD model has a slightly better performance than the BIR model. All the experiments using MDDR database shown that BIR model performs better than the BD model, whether with query fusion or non-query fusion approach. Meanwhile, it is found that query fusion results in poorer performance compared to non query fusion. For future work, it is suggested that the models to be tested on a wider range of activities using different bit strings. We also suggest that other probability-based model be explored in the retrieval of chemical compounds from chemical databases.

References

[1] Chen, V. Reynolds, C.H, "Performance of similarity measures in 2 D fragment-based similarity searching: comparison of structural descriptors and similarity coefficients", *Journal of Chemical Information Computer Sciences*,Vol 42,pp1407-1414, 2003.

[2] Willet , P., "Textual and chemical information processing: different domains but similar algorithms", Information Research, Vol 5 No 1, 1999.

[3] Cooper, W.S., "The formalism of probability theory in IT: A foundation or an encumbrance?" Proceeding of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Dublin Ireland: ACM, pp. 242-247, 1994.

[4] Efthimiadis, E.N "User Choices: A new yardstick for the evaluation of ranking algorithms for interactive query expansion", Information Processing & Management, Vol 31, No.4, pp605-620, 1995.

[5] Salton, G., McGill,M.J, *Introduction to modern Information Retrieval*, McGraw Hill Publishing Company, New York, 1983.

[6] Robertson,S.E., Spark Jones, "K. Relevance Weighting of Search Terms", *Journal of the American Society for International Science*, 27,pp.129-146, 1996.

[7] Shaw,W.M, "Term-Relevance Computation and Perfect Retrieval Performance", Information Processing & Management,Vol 31,No4,pp.491-498, 1995.

[8] Chow, C. and Liu, C. (1968). "Approximating discrete probability distributions with dependence trees."

IEEE Transactions on Information Theory, 14(3): pp. 462-467, 1968.

[9] Salim, N. and Godfrey, W.W.P. "Effectiveness of Probability Models for Compound Similarity Searching", *Journal of Advancing Information and Management Studies*, Special Issue on ICT in Health, Medicine and Biology, accepted.

[10] Salim, N, *Analysis and Comparison of Molecular Similarity Measures,* University of Sheffield: Ph. D Thesis, 2002.